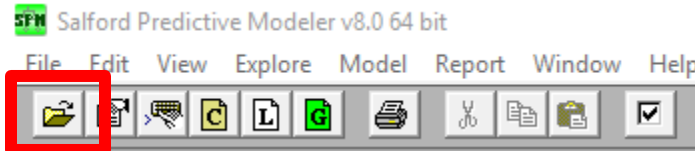


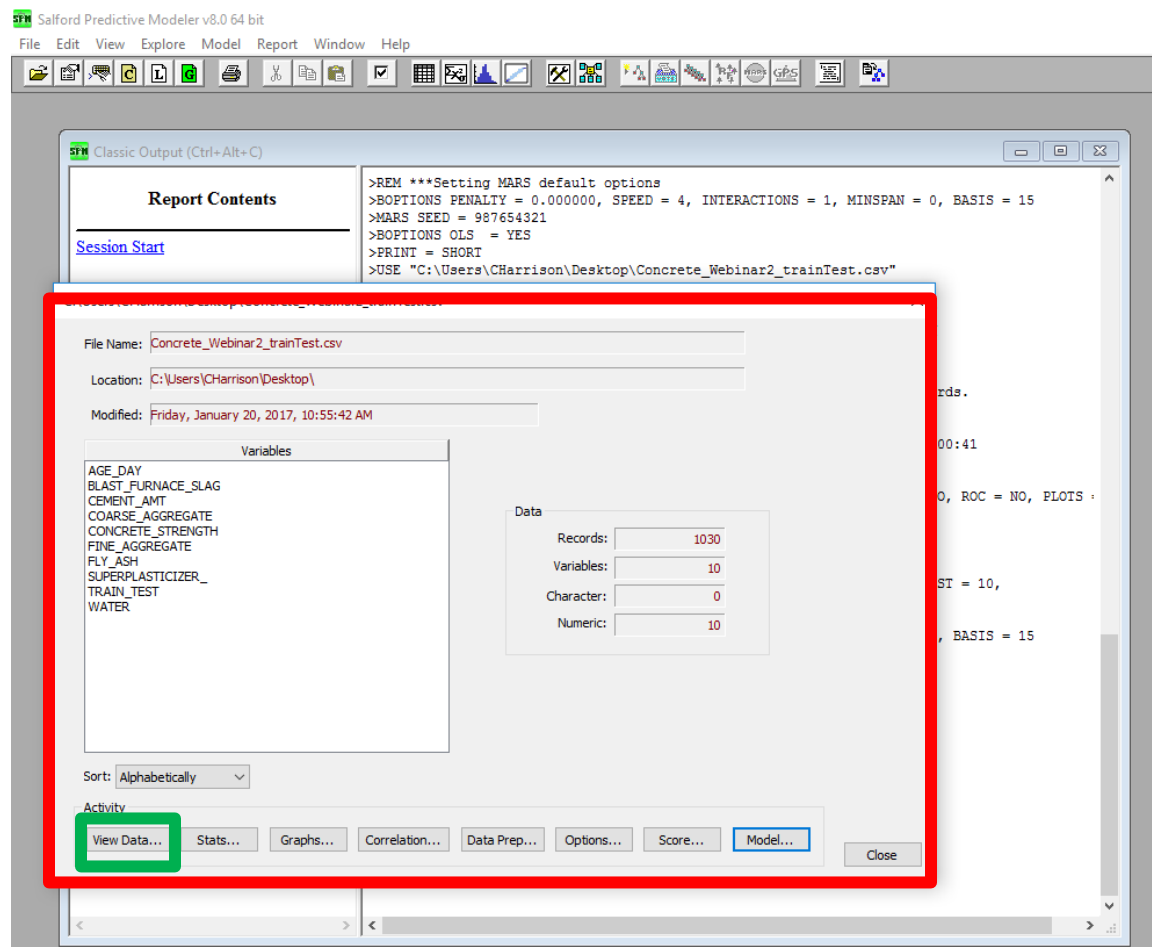
Step-by-Step Guide to CART® Software

Read in the data to SPM

1. Click the Open Data shortcut button



2. Double click on the data file name. After doing so you will see the **Activity Window**:



The Activity Window can be used to View the dataset, compute summary statistics, construct histograms, compute correlations and other measures of similarity, perform basic data preparation tasks (if desired), predict new observations, and build a model.

3. Click “View Data” (green rectangle above). The result will be the following:

Salford Predictive Modeler v8.0 64 bit

File Edit View Explore Model Report Window Help

Classic Output (Ctrl+Alt+C)

Report Contents

Session Start

>REM ***Setting MARS default options

C:\Users\CHarrison\Desktop\Concrete_Webinar2_trainTest.csv

	TRAIN_TEST	CEMENT_AMT	COARSE_AGG REGATE	BLAST_FURNA CE SLAG	FLY_ASH	WATER	SUPERPLASTI CIZER	FINE_AGGREG ATE	AGE_DAY	CONCRETE_S TRENGTH
1	0	540	1040	0	0	162	2.5	676	28	79.9861
2	0	540	1055	0	0	162	2.5	676	28	61.8874
3	0	332.5	932	142.5	0	228	0	594	270	40.2695
4	0	332.5	932	142.5	0	228	0	594	365	41.0528
5	0	198.6	978.4	132.4	0	192	0	825.5	360	44.2961
6	1	266	932	114	0	228	0	670	90	47.0298
7	1	380	932	95	0	228	0	594	365	43.6983
8	0	380	932	95	0	228	0	594	28	36.4478
9	1	266	932	114	0	228	0	670	28	45.8543
10	0	475	932	0	0	228	0	594	28	39.2898
11	0	198.6	978.4	132.4	0	192	0	825.5	90	38.0742
12	0	198.6	978.4	132.4	0	192	0	825.5	28	28.0217
13	0	427.5	932	47.5	0	228	0	594	270	43.013
14	0	190	932	190	0	228	0	670	90	42.3269
15	1	304	932	76	0	228	0	670	28	47.8138
16	0	380	932	0	0	228	0	670	90	52.9083
17	1	139.6	1047	209.4	0	192	0	806.9	90	39.358
18	0	342	932	38	0	228	0	670	365	56.142
19	0	380	932	95	0	228	0	594	90	40.5633
20	0	475	932	0	0	228	0	594	180	42.6206
21	0	427.5	932	47.5	0	228	0	594	180	41.8367
22	0	139.6	1047	209.4	0	192	0	806.9	28	28.2375
23	0	139.6	1047	209.4	0	192	0	806.9	3	8.06342
24	0	139.6	1047	209.4	0	192	0	806.9	180	44.2078
25	0	380	932	0	0	228	0	670	365	52.5167
26	0	380	932	0	0	228	0	670	270	53.3006
27	0	380	932	95	0	228	0	594	270	41.1514
28	0	342	932	38	0	228	0	670	180	52.1244
29	0	427.5	932	47.5	0	228	0	594	28	37.4275
30	0	475	932	0	0	228	0	594	7	38.6038
31	1	304	932	76	0	228	0	670	365	55.2601

Note: the variable TRAIN_TEST is used to distinguish between the LEARN data (i.e. the data used to build the tree; also called the “Training Data”) and the TEST data (i.e. the data used to validate the model; also called “Validation Data”)

4. I want to see if any variables are categorical, so I am going to compute summary statistics and look at the number of distinct levels for each variable. Note that any variables with character data (Example: “Male” or “Female”) are automatically read in as categorical variables but variables that have values 0 or 1 (Example: 0 is female and 1 is male) are read in as numeric. Click the Summary Statistics shortcut button below (red rectangle):



The result will be the following:

Descriptive Stats Setup

Variable Name	Include	Strata	Weight
AGE_DAY	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BLAST_FURNACE_SLAG	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CEMENT_AMT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN_TEST	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WATER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: **Alphabetically** Only Current Model Variables ☐ Select Vars ☐ Select ☐ Nest Strata ☐

☐ Fast Stats (No Tables, No Quantiles)
☒ Detailed Stats and Tables

Max. distinct values to track: All 1000
Max. distinct values to display: All 2000
Separate display for most and least common 5 values.

Filter: ☒ None ☐ Character ☐ Numeric ☐ Details to Classic Output

☐ Save to Grove

Dataset N Records: 1,030 Selected Variables: 10

Cancel OK

Click the “Include” label to highlight all variables and click “Select Vars” to select all variables

Descriptive Stats Setup

Variable Name	Include	Strata	Weight
AGE_DAY	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BLAST_FURNACE_SLAG	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CEMENT_AMT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN_TEST	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WATER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: **Alphabetically** Only Current Model Variables ☐ **Select Vars** ☐ Select ☐ Nest Strata ☐

☐ Fast Stats (No Tables, No Quantiles)
☒ Detailed Stats and Tables

Max. distinct values to track: All 1000
Max. distinct values to display: All 2000
Separate display for most and least common 5 values.

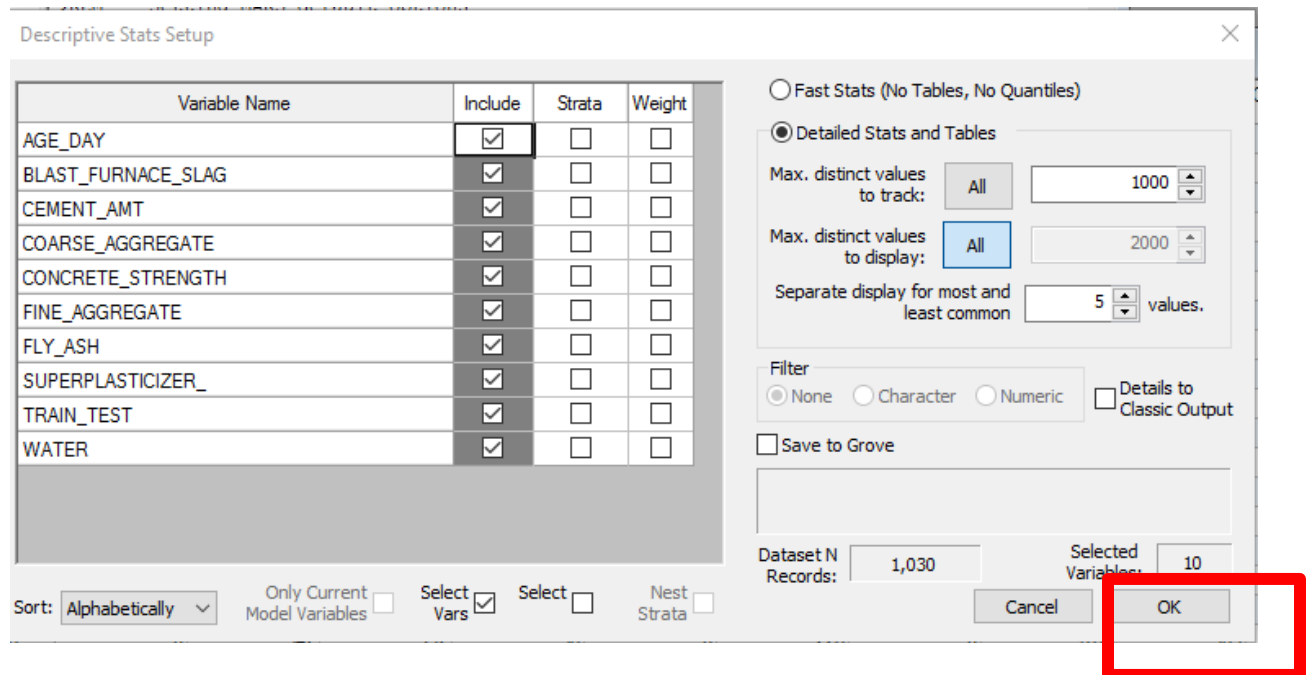
Filter: ☒ None ☐ Character ☐ Numeric ☐ Details to Classic Output

☐ Save to Grove

Dataset N Records: 1,030 Selected Variables: 10

Cancel OK

The result will be the following. Now click “OK”



The image shows the 'Descriptive Stats Setup' dialog box. On the left, a list of variables is shown with checkboxes for 'Include', 'Strata', and 'Weight'. All variables are checked under 'Include'. On the right, the 'Detailed Stats and Tables' radio button is selected. Below it, 'Max. distinct values to track' is set to 1000 and 'Max. distinct values to display' is set to 2000. 'Separate display for most and least common' is set to 5 values. The 'Filter' section has 'None' selected. 'Save to Grove' is unchecked. At the bottom, 'Dataset N Records' is 1,030 and 'Selected Variables' is 10. The 'OK' button is highlighted with a red rectangle.

Variable Name	Include	Strata	Weight
AGE_DAY	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BLAST_FURNACE_SLAG	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CEMENT_AMT	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN_TEST	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WATER	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: **Alphabetically** Only Current Model Variables ☐ Select Vars ☒ Select ☐ Nest Strata ☐

Fast Stats (No Tables, No Quantiles) ☐ Detailed Stats and Tables ☒

Max. distinct values to track: All 1000

Max. distinct values to display: All 2000

Separate display for most and least common 5 values.

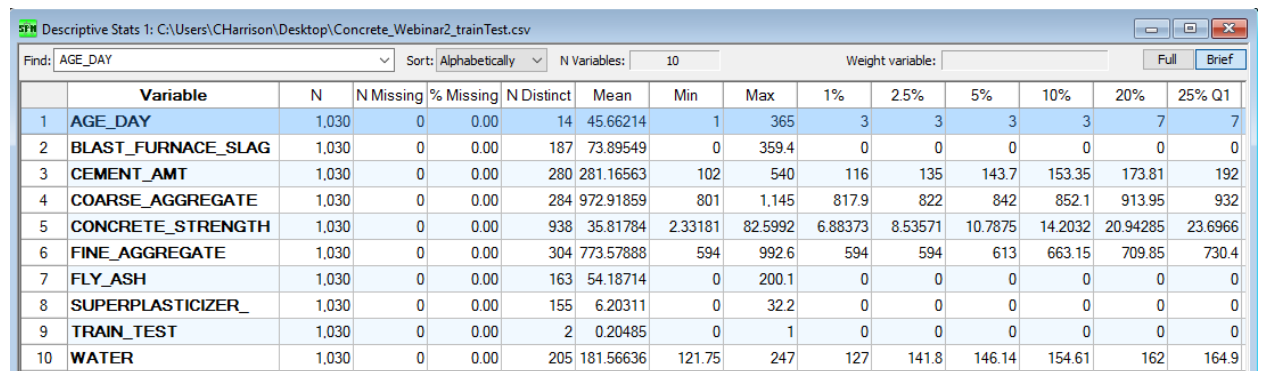
Filter: ☒ None ☐ Character ☐ Numeric ☐ Details to Classic Output

Save to Grove ☐

Dataset N Records: 1,030 Selected Variables: 10

Cancel OK

The result will be the following. Look at the N Distinct column. This column tells you how many unique values there are for each variable. Note that all variables have many values and are numeric so we can just treat them as continuous (TRAIN_TEST is NOT a predictor and is used to distinguish between the LEARN and TEST DATA).



The image shows the 'Descriptive Stats 1' window for the file 'C:\Users\CHarrison\Desktop\Concrete_Webinar2_trainTest.csv'. The window displays a table with 10 variables. The 'N Distinct' column shows the number of unique values for each variable. The 'Train Test' variable has only 2 distinct values.

Variable	N	N Missing	% Missing	N Distinct	Mean	Min	Max	1%	2.5%	5%	10%	20%	25% Q1
1 AGE_DAY	1,030	0	0.00	14	45.66214	1	365	3	3	3	3	7	7
2 BLAST_FURNACE_SLAG	1,030	0	0.00	187	73.89549	0	359.4	0	0	0	0	0	0
3 CEMENT_AMT	1,030	0	0.00	280	281.16563	102	540	116	135	143.7	153.35	173.81	192
4 COARSE_AGGREGATE	1,030	0	0.00	284	972.91859	801	1,145	817.9	822	842	852.1	913.95	932
5 CONCRETE_STRENGTH	1,030	0	0.00	938	35.81784	2.33181	82.5992	6.88373	8.53571	10.7875	14.2032	20.94285	23.6966
6 FINE_AGGREGATE	1,030	0	0.00	304	773.57888	594	992.6	594	594	613	663.15	709.85	730.4
7 FLY_ASH	1,030	0	0.00	163	54.18714	0	200.1	0	0	0	0	0	0
8 SUPERPLASTICIZER_	1,030	0	0.00	155	6.20311	0	32.2	0	0	0	0	0	0
9 TRAIN_TEST	1,030	0	0.00	2	0.20485	0	1	0	0	0	0	0	0
10 WATER	1,030	0	0.00	205	181.56636	121.75	247	127	141.8	146.14	154.61	162	164.9

Build a CART Model

1. Click the model setup shortcut button



The result will be the following:

Model Setup

Limits Costs Priors Penalty Lags Automate
Model Categorical Force Split Constraints Testing Select Cases Best Tree Method

Variable Selection

Variable Name	Target	Predictor	Categorical	Weight	Aux.
AGE_DAY	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BLAST_FURNACE_SLAG	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CEMENT_AMT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN TEST	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: Alphabetically

Filter: ☒ All/Selected ☐ Character ☐ Numeric

Target Type:
☒ Classification/Logistic Binary
☐ Regression
☐ Unsupervised

Set Focus Class...

Target Variable

Weight Variable

Number of Predictors: 10

Automatic Best Predictor Discovery:
☒ Off
☐ Discover only
☐ Discover and run

Maximum variables for each class: 8

After Building a Model: Save Grove...

Number of Predictors in Model: 10

Analysis Engine: CART Decision Tree

Cancel Continue Start

2. Setup the variables: Click the box in the Target column for the variable “CONCRETE_STRENGTH” (red rectangle below) and set the rest of the variables as predictors by clicking the “Predictor” label (green rectangle below) to highlight the column and then click “Select Predictors” checkbox (orange rectangle below). The result will be the following:

Model Setup

Limits Costs Priors Penalty Lags Automate
Model Categorical Force Split Constraints Testing Select Cases Best Tree Method

Variable Selection

Variable Name	Target	Predictor	Categorical	Weight	Aux.
AGE_DAY	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BLAST_FURNACE_SLAG	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CEMENT_AMT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN TEST	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: Alphabetically

Filter
☒ All/Selected ☐ Character ☐ Numeric

Target Type
☐ Classification/Logistic Binary
☒ Regression
☐ Unsupervised

Set Focus Class...

Target Variable
CONCRETE_STRENGTH

Weight Variable

Number of Predictors
9

Select Predictors ☒ Select Cat. ☐ Select Aux. ☐

Automatic Best Predictor Discovery
☒ Off
☐ Discover only
☐ Discover and run Maximum variables for each class 8

After Building a Model
Save Grove...

Number of Predictors in Model: 9

Analysis Engine
CART Decision Tree

Cancel Continue Start

3. Uncheck the checkbox in the “Predictor” column for the variable TRAIN_TEST because it is not a predictor variable (red rectangle below). Set the “Target Type” to be “Regression” (green rectangle below). The final result will be the following:

Model Setup

Limits Costs Priors Penalty Lags Automate
Model Categorical Force Split Constraints Testing Select Cases Best Tree Method

Variable Selection

Variable Name	Target	Predictor	Categorical	Weight	Aux.
CEMENT_AMT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN_TEST	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WATER	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: Alphabetically

Filter
☒ All/Selected ☐ Character ☐ Numeric

Select Predictors ☒ Select Cat. ☐ Select Aux. ☐

Target Type
☐ Classification/Logistic Binary
☒ Regression
☐ Unsupervised

Set Focus Class...

Target Variable
CONCRETE_STRENGTH

Weight Variable

Number of Predictors
8

Automatic Best Predictor Discovery
☒ Off
☐ Discover only
☐ Discover and run Maximum variables for each class 8

After Building a Model
Save Grove...

Number of Predictors in Model: 8

Analysis Engine
CART Decision Tree

Cancel Continue Start

4. Click the “Testing” Tab (red rectangle below) and click the radio button next to “Variable separates learn, test, (holdout):” **and** then click “TRAIN_TEST” (green rectangle below):

The screenshot shows the 'Model Setup' dialog box with the 'Testing' tab selected. The 'Testing' tab is highlighted with a red rectangle. The 'Variable separates learn, test, (holdout):' radio button is selected, and the 'TRAIN_TEST' option is highlighted in the list below it, both enclosed in a green rectangle. A purple rectangle highlights the 'Fraction of cases selected at random' and 'V-fold cross-validation' options.

Model Setup

Limits Costs Priors Penalty Loss Automate
Model Categorical Force Split Constraints **Testing** Select Cases Best Tree Method

Select Method for Testing

☐ No independent testing - exploratory model

☐ Fraction of cases selected at random: 0.20 ☒ Fast ☐ Exact

☐ Test sample contained in a separate file:

Cross-Validation

☐ V-fold cross-validation: Folds: 10 ☐ Save CV models to grove

☐ Save OOB Predictions:

☐ Variable determines CV bins:

☒ Variable separates learn, test, (holdout): TRAIN_TEST

TRAIN_TEST

Automatic Best Predictor Discovery

☒ Off

☐ Discover only

☐ Discover and run Maximum variables for each class: 8

After Building a Model

Save Grove...

Number of Predictors in Model: 8

Analysis Engine

CART Decision Tree

Cancel Continue Start

This setting ensures that the LEARN and TEST data are the same for CART and Random Forest so we can compare their performance on the TEST data (we really don't care about the LEARN performance). Now if you want to use a test sample but you aren't worried about comparisons then select either cross validation (for smaller and medium sized datasets), "Fraction of cases selected at random", or "Test sample contained in a separate file"(purple rectangle above)

5. Click the “Method” tab

The screenshot shows the 'Model Setup' dialog box with the 'Method' tab selected. The 'Classification Trees' section has 'Gini' selected. The 'Regression Trees' section has 'Least Squares' selected. The 'Start' button is highlighted with a red rectangle.

Model Setup

Method

Select Splitting Method

Classification Trees

- ☒ Gini
- ☐ Symmetric Gini
- ☐ Entropy
- ☐ Class Probability
- ☐ Twoing
- ☐ Ordered Twoing
- ☐ Differential Lift

Favor Even Splits

Less More

Regression Trees

- ☒ Least Squares
- ☐ Least Absolute Deviation

☐ Use Linear Combinations for Splitting

Minimum node sample size for linear combinations:

Variable deletion significance level:

Number of nodes likely to be split by linear combinations in maximal tree: ☐ Automatic

☐ Create Advanced Variable Lists

Automatic Best Predictor Discovery

- ☒ Off
- ☐ Discover only
- ☐ Discover and run

Maximum variables for each class:

After Building a Model

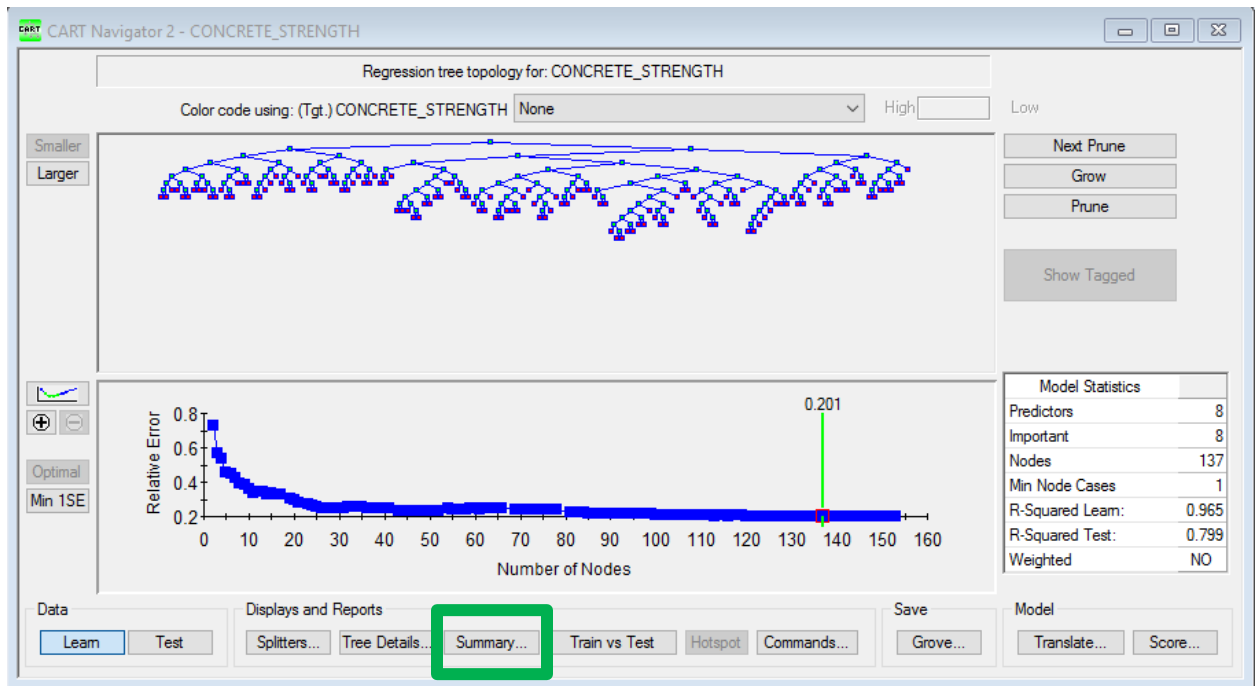
Number of Predictors in Model:

Analysis Engine

Note: the method for regression trees is set to “Least Squares” by default. If you want to change the method for either regression or classification trees then you can do so in this tab. I just want to show you this, so just leave the default settings for now.

Click “Start” to build the CART model (red rectangle above).

6. Here is the optimal CART tree. To view the summary statistics for this model click “Summary...” (green rectangle below)

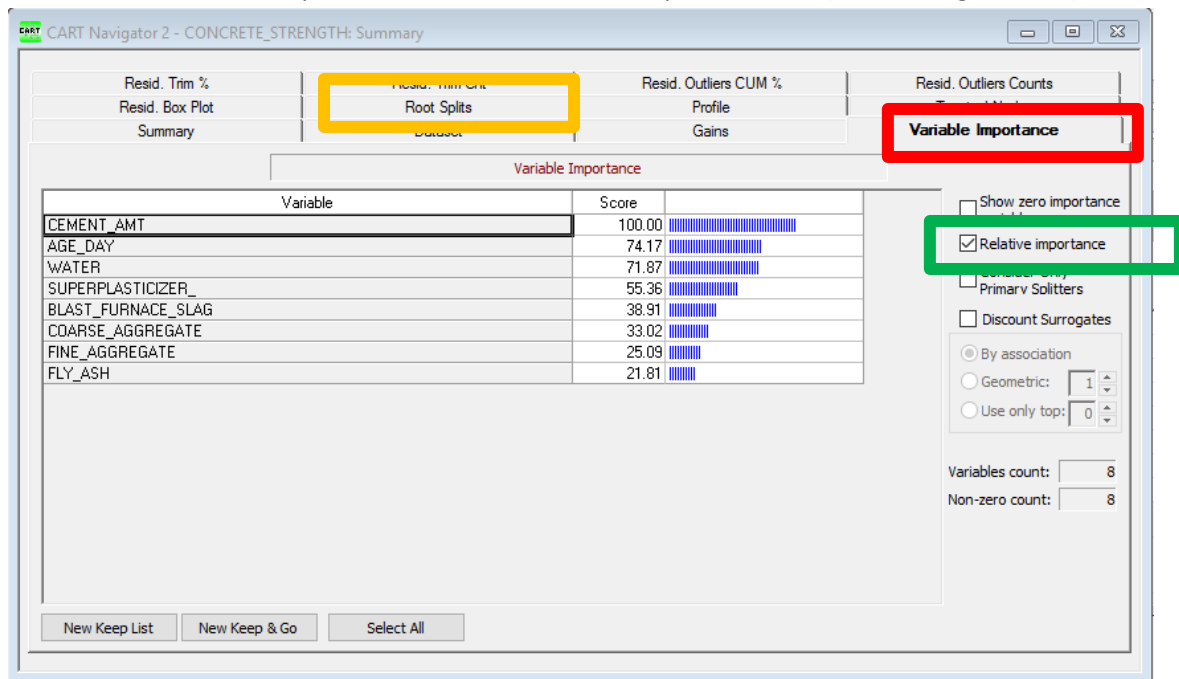


Here are the results. Note the value for the MSE on the test data is 55.99:

The screenshot shows the CART Navigator 2 - CONCRETE_STRENGTH: Summary window. The 'Model Summary' section displays various error measures for the model. The 'MSE' row is highlighted with a green rectangle, showing a value of 55.99059 for the Test data.

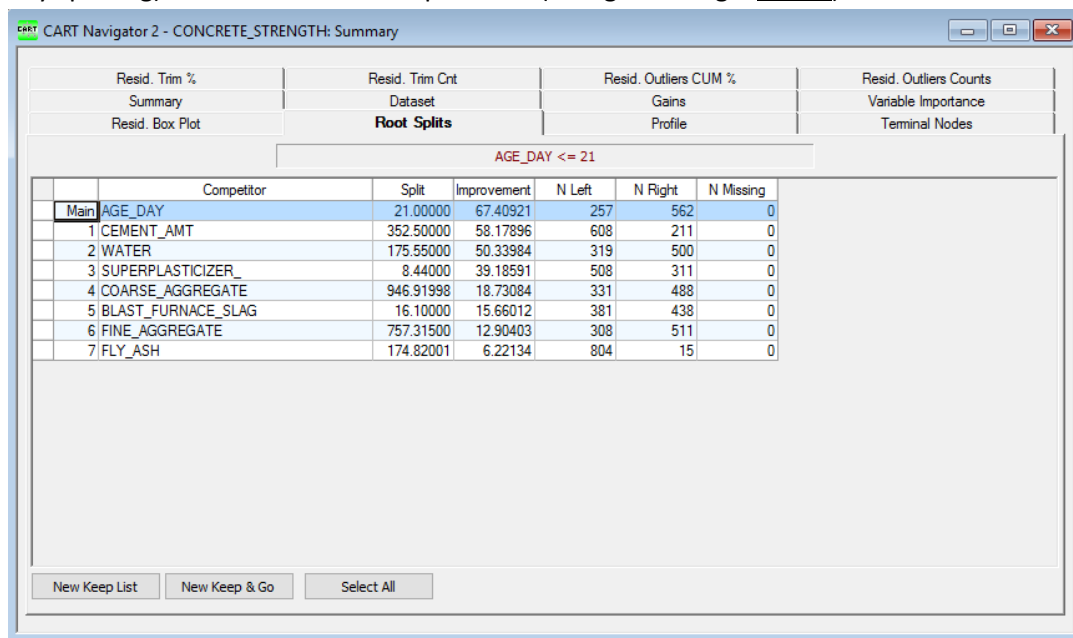
Name	Learn	Test
MSE	9.63732	55.99059
MAE	2.22332	3.23213
MAPE	0.07803	0.18648
SSY	228,170.06727	58,788.06893
SSE	7,892.96272	11,814.01400
R-Sq	0.96541	0.79904
R-Sq Norm	0.96541	0.80312
AIC	1,871.56142	865.31374
AICc	1,871.73920	866.02661
BIC	1,909.22610	892.12860
Relative Error	0.03459	0.20096

7. To view the “Variable Importance” click the variable importance tab (red rectangle below)



Note: currently Relative Importance is currently checked so each raw importance score has been divided by the maximum score which results in the most important variable having a value of 100% and so on. To view the raw importance score simply uncheck the checkbox next to “Relative Importance” (green rectangle above)

8. To view the improvement scores in the root node (i.e. the node at the top of the tree prior to any splitting) then click the “Root Splits” tab (orange rectangle above).



Note: AGE_DAY and CEMENT_AMT both have high improvement scores which may indicate (not always) that they are highly correlated.

Viewing detailed summary statistics for the 1 Standard Error CART Tree

9. To view detailed results (i.e. MSE, SSE, etc.) for the minimum 1 standard error tree do the following:
 - a) Setup the model as described previously
 - b) Go the “Best Tree” tab (red rectangle below) and click the radio button next to “Within one standard error minimum” (green rectangle below). Now click “Start” (orange rectangle below)

The screenshot shows the 'Model Setup' dialog box with the 'Best Tree' tab selected. The 'Standard Error Rule' section has the 'Within one standard error of minimum' option selected. The 'Start' button is highlighted in orange.

Model Setup

Limits Costs Priors Penalty Lags Automate
Model Categorical Force Split Constraints Testing Select Case **Best Tree** Method

Parameters Influencing Selection of Best Tree

Standard Error Rule

☐ Minimum cost tree regardless of size

☒ Within one standard error of minimum

☐ Set S.E. rule = 1

Variable Importance Formula

☒ All surrogates count equally

☐ Discount surrogates

Weight = 1

Surrogates And Competitors

Number of surrogates to use for constructing tree: 5

Number of competitors to track: 5

Recall Defaults

Automatic Best Predictor Discovery

☒ Off

☐ Discover only

☐ Discover and run

Maximum variables for each class: 8

After Building a Model

Save Grove...

Number of Predictors in Model: 8

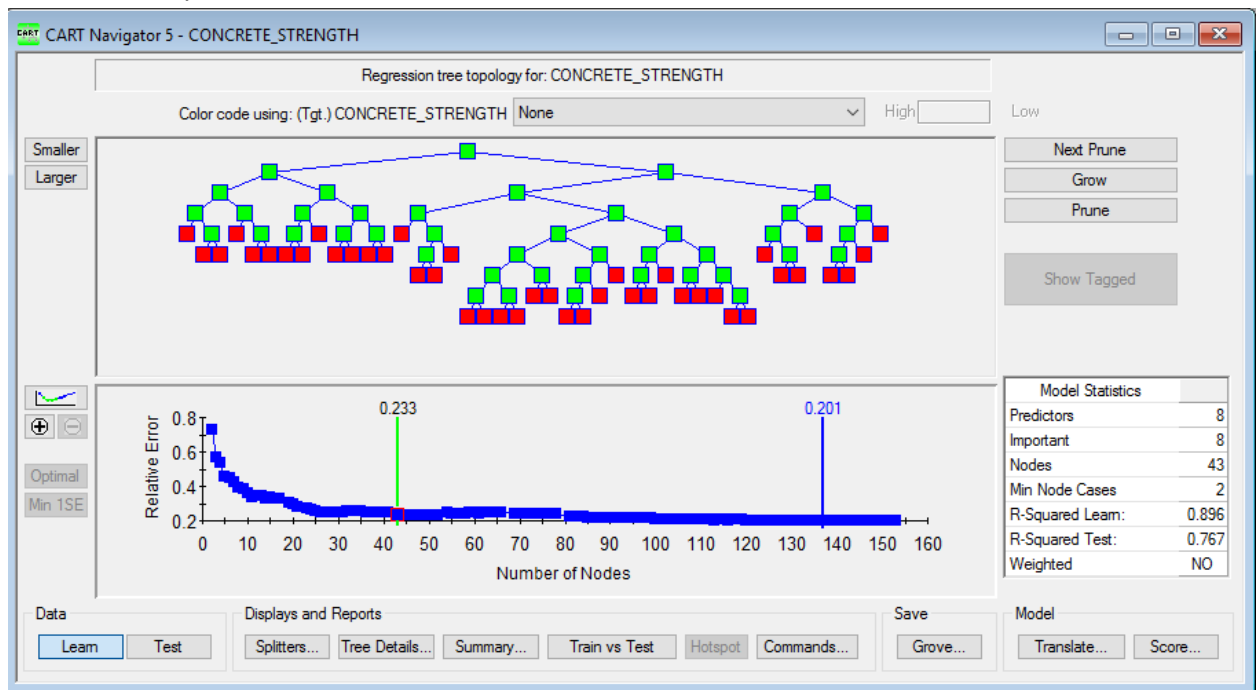
Analysis Engine

CART Decision Tree

Cancel Continue **Start**

The tree that you see will be the 1 standard error tree.

10. Click "Summary"



Here are the results. Note that you can now see detailed summary statistics for this tree:

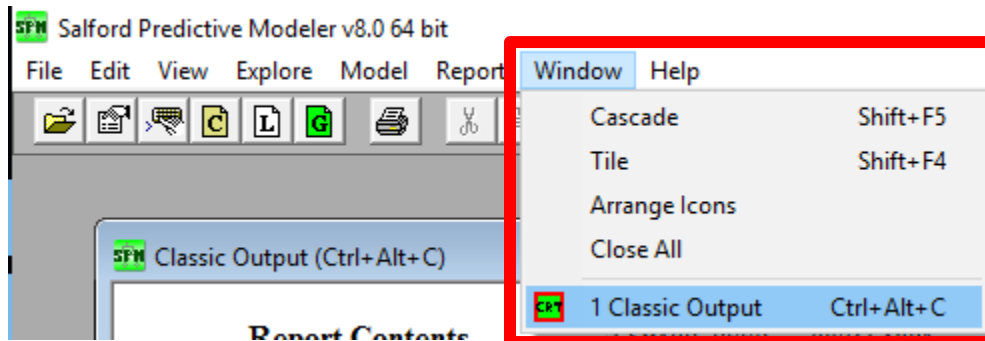
The screenshot shows the 'CART Navigator 5 - CONCRETE_STRENGTH: Summary' window. It displays detailed model statistics. On the left, there are tabs for 'Resid. Trim %', 'Resid. Box Plot', and 'Summary' (selected). Below these are fields for 'Model' (Target: CONCRETE_STRENGTH, Total N: 1,030, Wgt Total N: 1030.00, N Cat: Regression, Predictors: 8, Focus Class:). The main area is titled 'Model Summary' and contains a table of 'Model error measures'.

Name	Learn	Test
RMSE	5.38578	8.06545
MSE	29.00665	65.05151
MAD	4.26332	6.16996
MAPE	0.16194	0.23672
SSY	228,170.06727	58,788.06893
SSE	23,756.44550	13,725.86945
R-Sq	0.89588	0.76652
R-Sq Norm	0.89588	0.76974
AIC	2,774.00304	896.96287
AICc	2,774.18082	897.67574
BIC	2,811.66771	923.77774
Relative Error	0.10412	0.23348

At the bottom, there are buttons for 'Commands...', 'Translate...', 'Score...', and 'Save Grove...'.

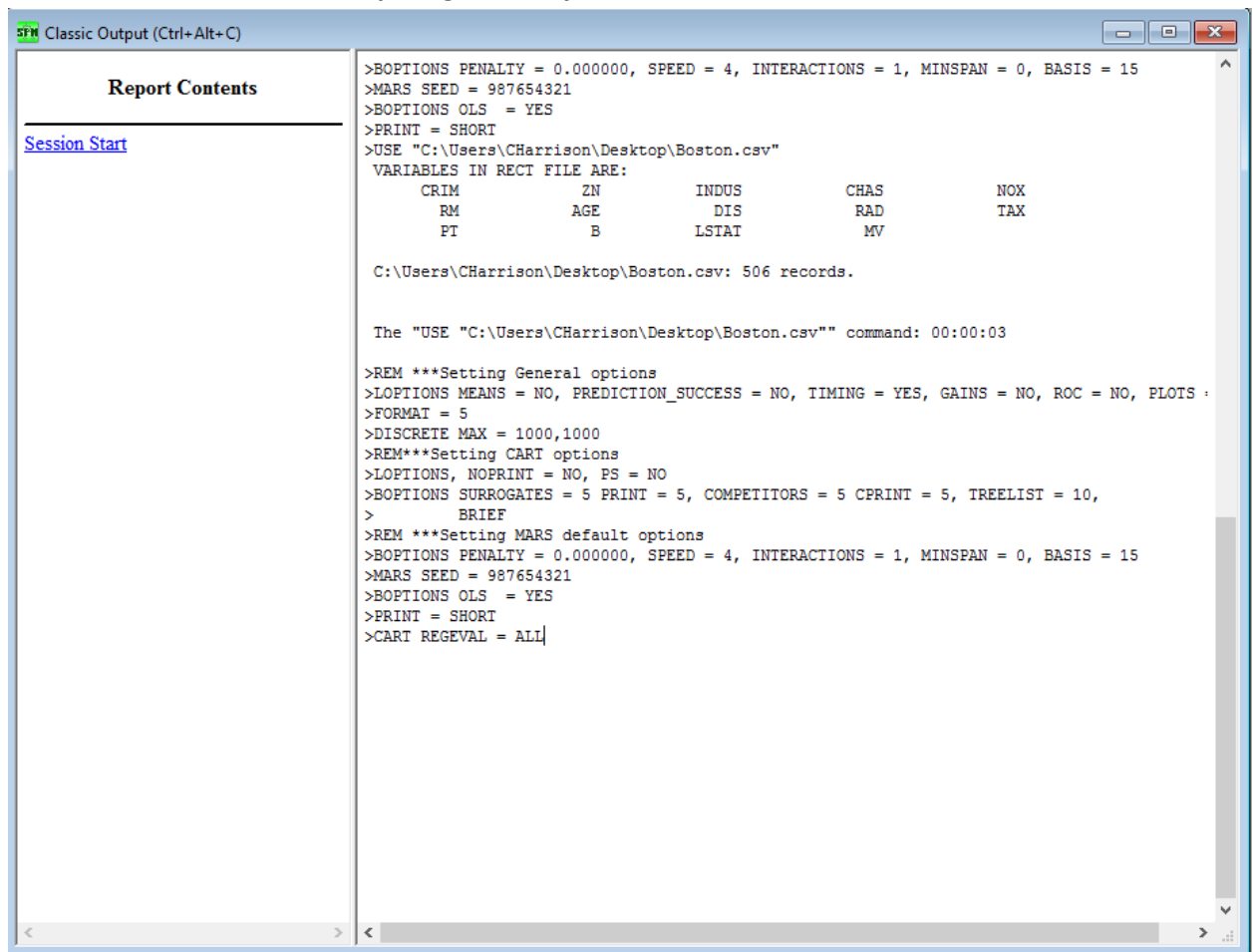
Viewing detailed summary statistics for all trees in the pruning sequence

1. Go to the Classic Output Window by clicking Window > Classic Output



You will see the following window. Now type `CART REGEVAL = ALL`

Press Enter to submit the command. Now build a CART model and now you can see summary statistics for each tree in the pruning sequence. **Note: this will result in a longer computing time because now we are computing summary statistics for dozens of trees.**

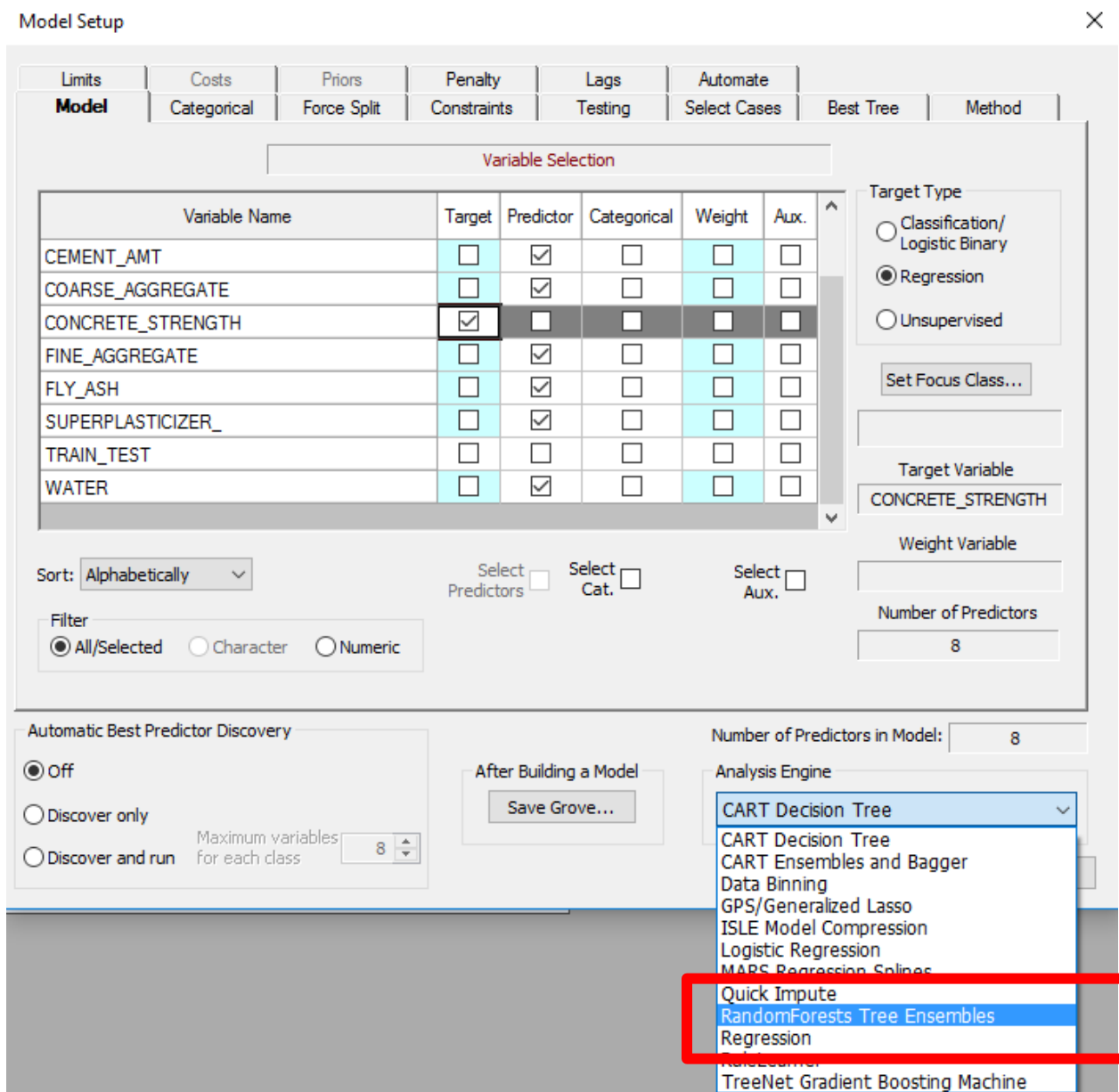


Building a Random Forest Model

1. Click the model setup shortcut button



2. Set the analysis engine to RandomForests Tree Ensembles (Red Rectangle below)



3. Since we want to compare the Random Forest with CART tree and the linear regression model we will use the same LEARN and TEST partition as was used for those models. Click the “Testing” Tab (red rectangle below) and click the radio button next to “Variable separates learn, test, (holdout):” **and** then click “TRAIN_TEST” (green rectangle below):

Model Setup

Class Weights | Lags | Automate | Select Cases | Random Forests

Model | Categorical | **Testing**

Select Method for Testing

☐ Out of bag data used for testing

☐ Fraction of cases selected at random: 0.20 ☒ Fast ☐ Exact

☐ Test sample contained in a separate file:

Cross-Validation

☐ V-fold cross-validation: Folds: 10 ☐ Save CV models to grove

☐ Save OOB Predictions:

☐ Variable determines CV bins:

☒ Variable separates learn, test, (holdout): TRAIN_TEST

Automatic Best Predictor Discovery

☒ Off

☐ Discover only

☐ Discover and run Maximum variables for each class: 8

After Building a Model

Save Grove...

Number of Predictors in Model: 8

Analysis Engine

Random Forests Tree Ensembles

Cancel Continue Start

4. Click the Random Forests tab (red rectangle below). Set the number of trees to 400 (green rectangle below) and uncheck the checkbox next to “Create Full Proximity Matrix” (orange rectangle below). Click “Start” (purple rectangle below) to build a Random Forest.

The screenshot shows the 'Model Setup' dialog box with the 'Random Forests' tab selected. The 'Number of trees to build' is set to 400. The 'Create Full Proximity Matrix' checkbox is unchecked. The 'Start' button is highlighted.

Model Setup

Random Forests Options

Options

Number of trees to build: 400

N predictors: Exactly 3

Frequency of progress reports: 10

Number of proximal cases to track (0 to disable): AUTO

Bootstrap sample size: AUTO

Parent node minimum cases Regression Recommended Min: 5

Defaults

☐ Save Results to Files

☒ Parallel Coordinates

☒ Outliers And Scaling Dimensions

☒ Probabilities And Class Predictions

☒ Partial Proximity ☒ Full Proximity

☒ Imputed ☒ Prototypes

☐ Create Full Proximity Matrix

If the number of records is less than or equal to: 10000

Post-processing

☒ Suppress all post-processing for Classification models

☐ Use advanced missing value imputation

Automatic Best Predictor Discovery

☒ Off

☐ Discover only

☐ Discover and run

Maximum variables for each class: 8

After Building a Model

Save Grove...

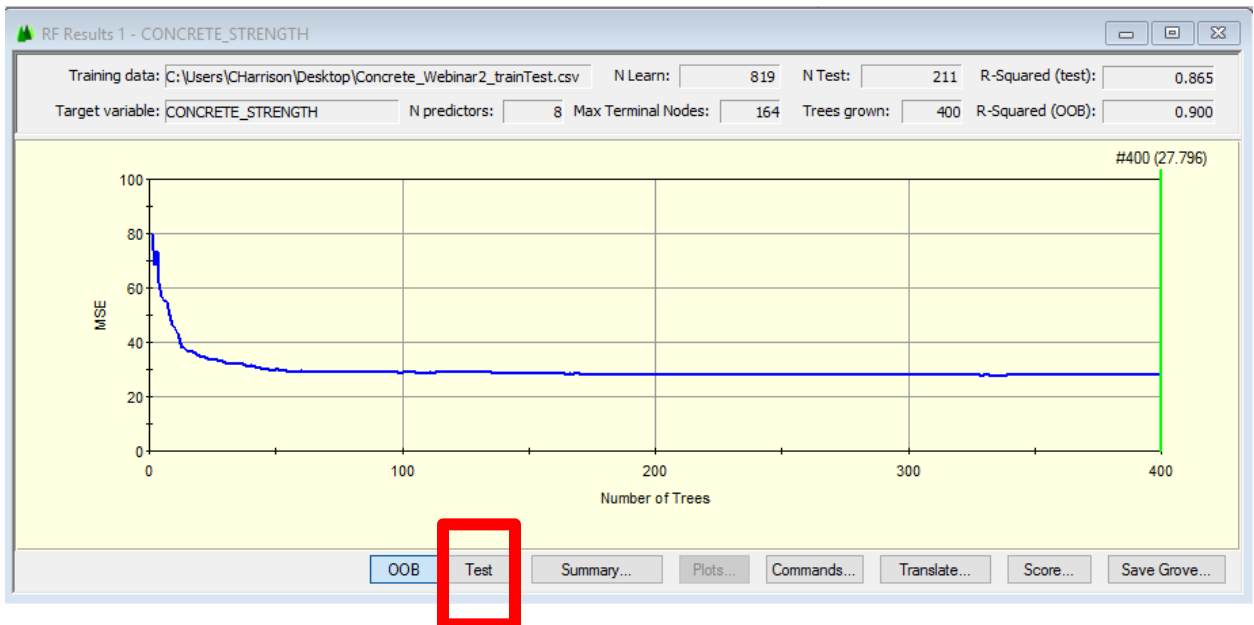
Number of Predictors in Model: 8

Analysis Engine

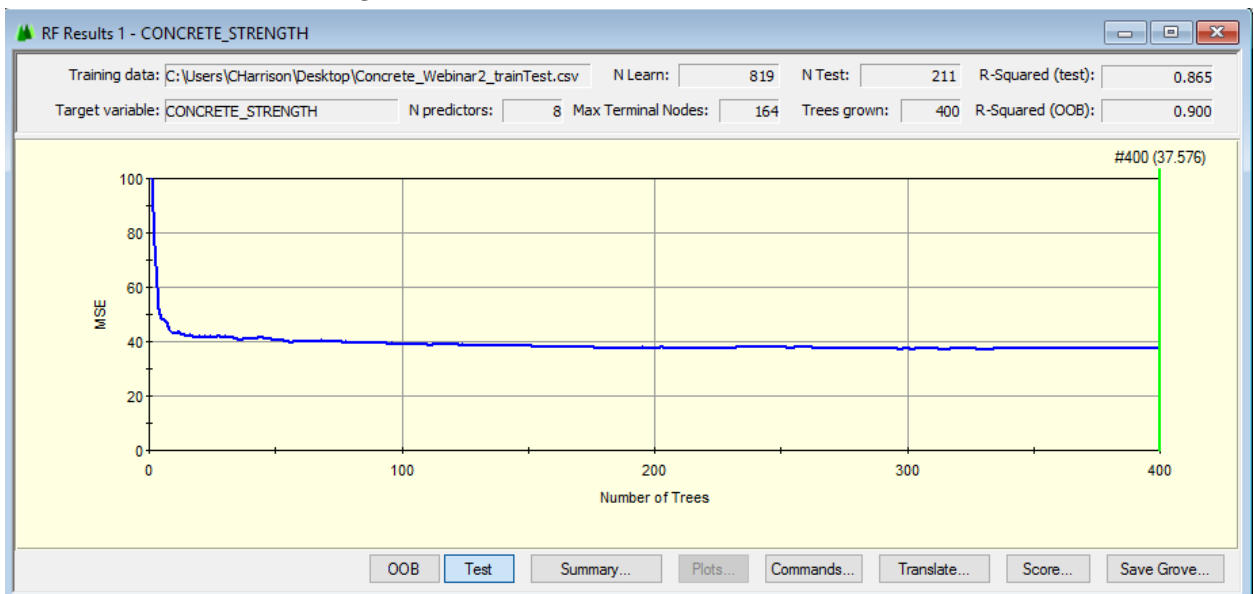
RandomForests Tree Ensembles

Cancel Continue **Start**

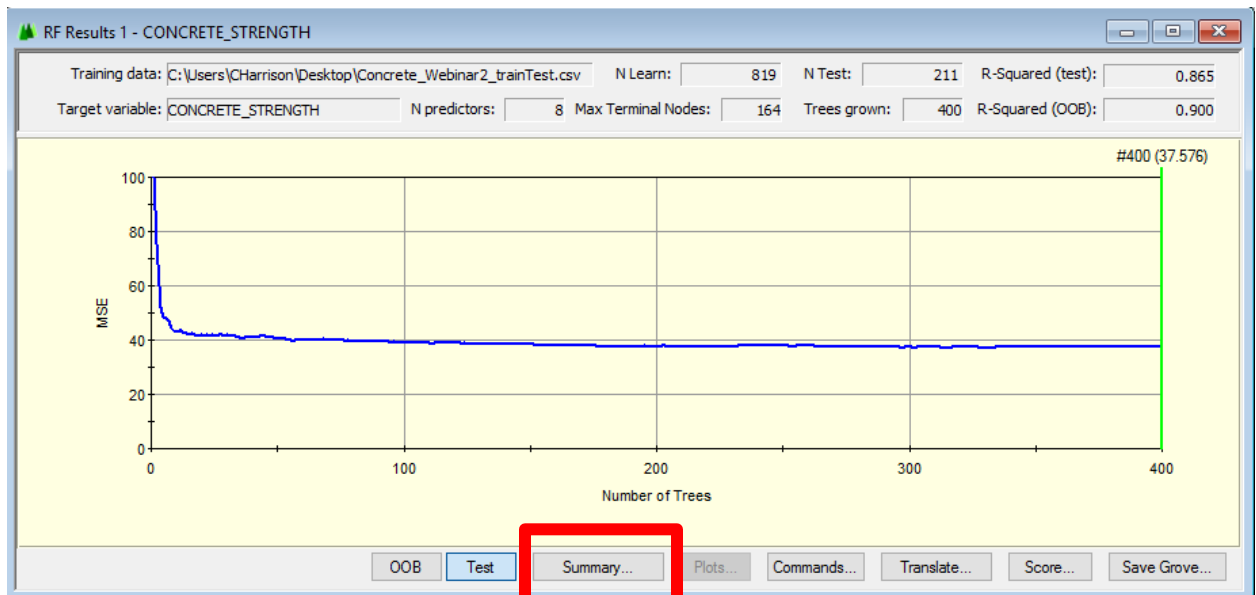
Note: the first error measurement that you will see is the “OOB” error (OOB = “Out Of Bag”), but since we want to compare our RandomForest model directly with the CART and linear regression model we need to click “Test” to view the error for the test data that we defined earlier in the Testing tab. Click “Test” (red rectangle below)



The result will be the following:



5. Now click “Summary” (red rectangle below) to view the results of the RandomForests. The model statistics that you see correspond to the optimal number of trees which in this case is 400



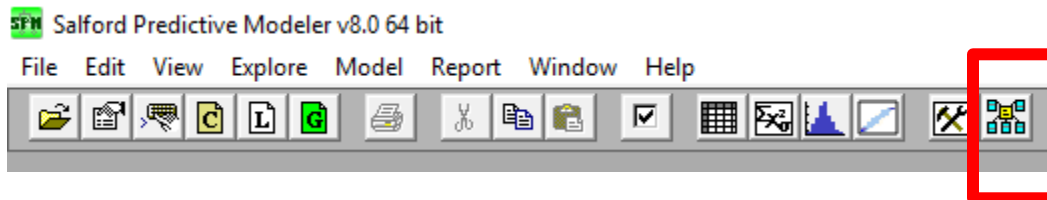
The results will be the following. Pay attention to the model statistics in the “Test” Column (red rectangle below):

RF Results 1 - CONCRETE_STRENGTH: Summary for 400 Committee Trees

Resid. Trim Cnt	Resid. Outliers CUM %	Resid. Outliers Counts	Resid. Box Plot
Summary	Dataset	Variable Importance	Resid. Trim %
Model Summary			
Model			
Target: CONCRETE_STRENGTH			
Total N: 1,030			
Wgt Total N: 1030.00			
N Cat: Regression			
Predictors: 8			
Focus Class:			
Model error measures			
Name	OOB	Test	
RMSE	5.27220	6.12991	
MSE	27.79612	37.57577	
MAD	3.94957	4.37378	
MAPE	0.14660	0.17092	
SSY	228,170.06727	58,788.06893	
SSE	22,765.02380	7,928.48846	
R-Sq	0.90023	0.86513	
R-Sq Norm	0.91410	0.87835	

Commands... Translate... Score... Save Grove...

6. Now we are going to determine the optimal number of variables to randomly select at each split in CART tree in the Random Forest. Click the model setup button again to open up the Random Forest model interface:



The result will be the following. Click “Automate” (red rectangle below)

Model Setup

Class Weights | Lags | **Automate** | Select Cases | Random Forests

Model | Categorical | Testing

Variable Selection

Variable Name	Target	Predictor	Categorical	Weight
AGE_DAY	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BLAST_FURNACE_SLAG	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CEMENT_AMT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
COARSE_AGGREGATE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONCRETE_STRENGTH	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FINE_AGGREGATE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FLY_ASH	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUPERPLASTICIZER_	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TRAIN TEST	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sort: Alphabetically

Filter: ☒ All/Selected ☐ Character ☐ Numeric

Target Type: ☐ Classification/Logistic Binary ☒ Regression ☐ Unsupervised

Set Focus Class...

Target Variable: CONCRETE_STRENGTH

Weight Variable:

Number of Predictors: 8

Automatic Best Predictor Discovery: ☒ Off ☐ Discover only ☐ Discover and run

Maximum variables for each class: 8

After Building a Model: Save Grove...

Number of Predictors in Model: 8

Analysis Engine: RandomForest Tree Ensembles

Cancel Continue Start

7. Click on “RFNPREDs” (red rectangle below) and then click “Add” (green rectangle below)

Model Setup

Model | Categorical | Testing | Select Cases | Random Forests

Class Weights | Lags | **Automate**

Automate Types

- ATOM
- DATASHIFT
- DRAW
- EVERYTHING
- FLIP
- KEEP
- LEARN CURVE
- LOVO
- MODELS
- ONEOFF
- RFNPREDs**
- SHAVING RFE

Vary the number of potential splitters at a node in Random Forests.

☐ Generate Individual Model Reports

Selected Automates

Add

Remove

Remove All

Automate Options

Option	Value
--------	-------

Set Automate Options

Automatic Best Predictor Discovery

☒ Off

☐ Discover only

☐ Discover and run

Maximum variables for each class: 8

After Building a Model

Save Grove...

Number of Predictors in Model: 8

Analysis Engine

RandomForests Tree Ensembles

Cancel Continue Start

The result will be the following (Note: as a shortcut you can just double click “RFNPRED5”). Click into the box next to “Values” (green rectangle below) and type the numbers (with commas) 1,2,3,4,5,6,7,8. This will construct 8 Random Forests: the first is built by selecting one variable randomly at each split in each tree in the forest, the second is built by selecting two variables randomly at each split in each tree in the forest, and so on. Note: we select these variables randomly, determine the best split for each of these selected variables and then the final split is the variable and split combination that most reduces the model error.

Model Setup

Model | Categorical | Testing | Select Cases | Random Forests

Class Weights | Lags | **Automate**

Automate Types

- ATOM
- DATASHIFT
- DRAW
- EVERYTHING
- FLIP
- KEEP
- LEARN CURVE
- LOVO
- MODELS
- ONEOFF
- PARTITION
- SEED**
- SHAVING RFE
- STEPWISE

Add | Remove | Remove All

Builds models with identical configuration using different Random Number Seed each time.

Selected Automates

- RFNPRED5**

Vary the number of potential splitters at a node in Random Forests.

Automate Options

Option	Value
RFNPRED5	
Values	1,2,3,4,5,6,7,8
Defaults	EIGHTHSQR,QUARTILE,HALFSQR,SQR,SQRX2
N Predictors < 64	1, 2, 4, 7, 11

Set Automate Options

☐ Generate Individual Model Reports

Automatic Best Predictor Discovery

☒ Off

☐ Discover only

☐ Discover and run

Maximum variables for each class: 8

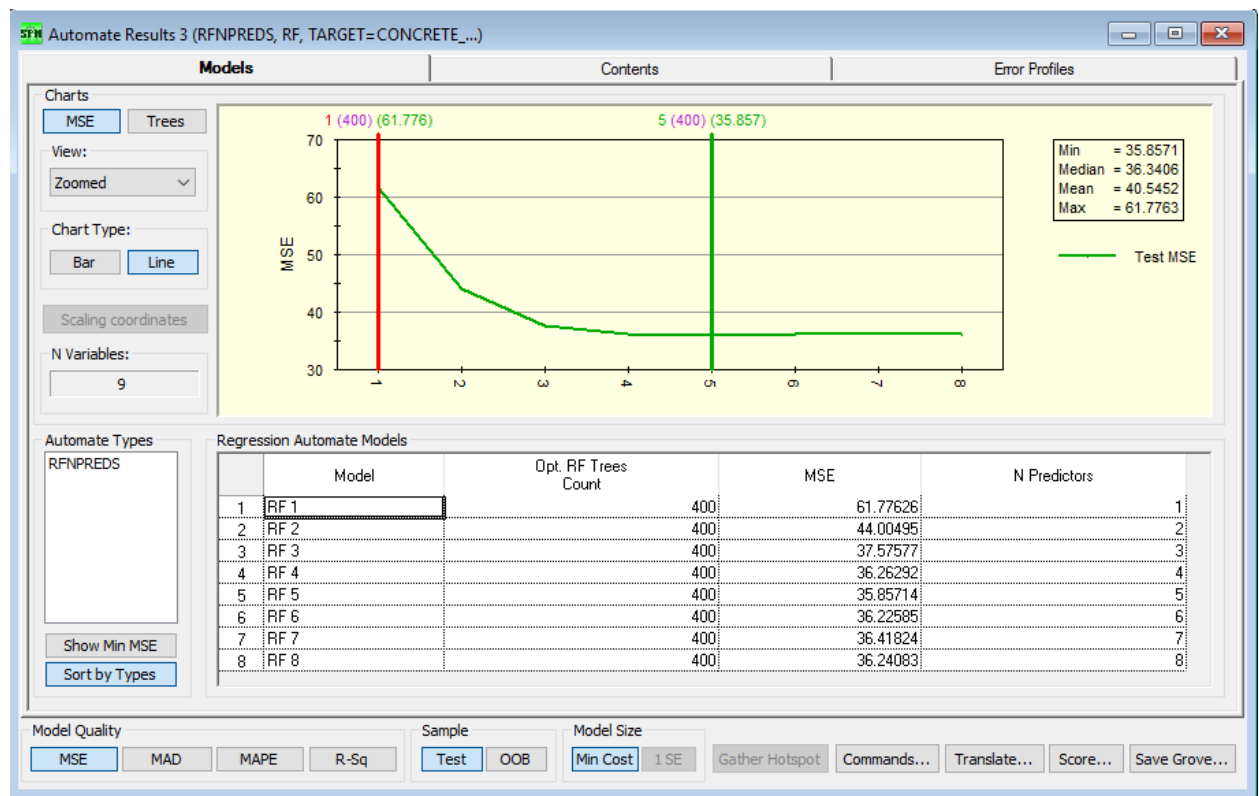
After Building a Model: Save Grove...

Number of Predictors in Model: 8

Analysis Engine: Random Forests Tree Ensembles

Cancel | Continue | Start

The result of Automate RFNPREDs is the following:



The optimal number of variables to randomly choose at each split is 5. Note: you can view any of the 8 Random Forests that were just constructed by double clicking "RF1", "RF2", and so on. Here is what you will see if you double click "RF3": it is the same as if you built this model by itself.

